

**EUROPEAN PATENT APPLICATION**

(43) Date of publication:  
04.03.1998 Bulletin 1998/10

(51) Int Cl.<sup>6</sup>: **H04L 12/56**

(21) Application number: **97480049.2**

(22) Date of filing: **19.08.1997**

(84) Designated Contracting States:  
**AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC  
NL PT SE**  
Designated Extension States:  
**AL LT LV RO SI**

- Onvural, Raif O.  
Cary, NC 27513 (US)
- Huynh, Lap T.  
Apex, NC 27502 (US)
- Gun, Levent  
Hopkinton, MA 01748 (US)
- Van Vu, Ken  
Cary, NC 27513 (US)
- Ling, Bouchung  
Cary, NC 27513 (US)

(30) Priority: **22.08.1996 US 701624**

(71) Applicant: **INTERNATIONAL BUSINESS  
MACHINES CORPORATION**  
Armonk, NY 10504 (US)

(74) Representative: **de Pena, Alain**  
**Compagnie IBM France**  
**Département de Propriété Intellectuelle**  
**06610 La Gaude (FR)**

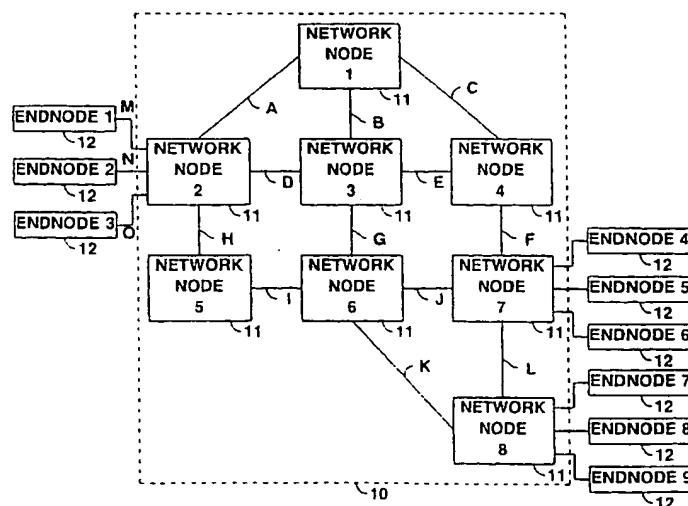
(72) Inventors:  
• Marin, Gerald Arnold  
Chapel Hill, NC 27514 (US)

(54) **Adaptive rate-based congestion control in packet networks**

(57) An enhanced adaptive rate-based congestion control system for packet transmission networks uses the absolute rather than the relative network queuing delay as the measure of congestion in the network. Other features of the congestion control system include test transmissions only after a predetermined minimum time, after the receipt of an acknowledgment from the previous test, or transmission of a minimum data burst,

whichever takes longest. The congestion control system also provides a small reduction in rate at low rates and a large reduction in rates at high rates. A logarithmic rate control function provides this capability. Rate damping is provided by changing all of the values in a rate look-up tables in response to excessive rate variations. Finally, the fair share of the available bandwidth is used as the starting point for rates at start-up or when a pre-defined rate damping region is exited.

**FIG. 1**



gestion control mechanism is provided which overcomes the problems of the prior art systems to provide enhanced congestion control in the face of rapid changes in network usage, using absolute rather than relative measures of current congestion and providing admission functions more responsive to the state of congestion in the network. More particularly, the congestion control system of the present invention provides a spacing function for determining the minimum interval between launching of two back-to-back cells entering the network. This spacing function is, of course, based on the currently granted transmission rate.

Furthermore, a sampling policy is enforced in which a sample packet is not set unless a predetermined minimum sampling period has expired, an acknowledgment of the previous sample packet has been received, and a specified amount of data (called the sampling burst) has been transmitted since the last sampling packet. This sampling policy ensures minimum sampling overhead since the overhead is dominated by the sampling burst when the round trip delay is short (ensuring maximum data throughput) and is dominated by the round trip delay when that delay is long (ensuring minimum sample frequency).

Like the prior art systems, the delay estimation of the present invention is based on the inter-arrival times of successive sample packets. The congestion, measured as network queuing delay, is a function of the difference in departure times of successive sampling packets (calculated from the time stamps in these packets) and the difference in arrival times of these successive sampling packets (measured at the receiving station), assuming that the previous network queuing delay is known. Assuming that the network delay is initially zero, the path delay quickly reduces to the minimum delay suffered by any sample packet. This value can then be used throughout the connection as a close estimate of the actual minimum path delay, allowing calculation of the actual current path delay rather than relying only on relative path delay changes. Using these actual current path delay values, connections are allowed to increase their rates when no incremental delay is detected, to maintain their current rates when the desired network utilization is achieved, to take a small reduction from their current rates when a slight congestion is detected, and to take a large reduction from their current rates when a serious congestion is detected.

Finally, a rate change policy is followed which maximizes efficiency, stability and fairness by providing a large increase in the transmission rate when that rate is low, thereby promoting efficiency, but by providing a small increase in transmission rate when that rate is high, thereby promoting stability. Similarly, reductions in the transmission rate can also be high when the rate is low and low when the rate is high. In accordance with the present invention, a logarithmic function is used to control changes in the transmission rate in response to changes in congestion. This strategy is modified in response to the even sharing of the remaining bandwidth to make the sharing of the overall bandwidth fair.

More particularly, the above-described rate change policy is implemented by a table-driven technique where the permissible rates for each source are stored in a table with the desired logarithmic value changes in successive table entries. Acknowledgments (ACKs), decrease ACKs and increase ACKs, step up or down through the entries in the table to access the next admissions rate. As previously noted, the table entries implement the policy of providing large changes in rate when the rate is low and small changes in rate when the rate is high. In addition, a rate damping region is defined within which it is assumed that even and fair sharing occurs. Although the damping region is initially defined to bracket the starting rate, this damping region is moved up and down in the table in response to movements outside of the damping region. At the same time, the table values are increased or decreased multiplicatively by a preselected factor. This procedure increases fairness over connections that may have different starting times by favoring connections having lower table values compared to connections having higher table values.

In accordance with another feature of the invention, in order to further improve fairness, the enhanced adaptive rate-based congestion control system of the present invention utilizes the current available bandwidth on the network to compute an initial fair share allocation for each new non-reserved connection as that connection is set up. This computation avoids the long start up time during which the prior art new connections must adapt to the current state of the network.

A major advantage of the present invention is that it is an end-to-end control system, requiring no special processing at intermediate nodes along the route, and thus requiring no changes in the processing done at the intermediate nodes.

### **Brief Description of the Drawings**

FIG. 1 shows a general block diagram of a packet communications network in which enhanced rate-based adaptive congestion control in accordance with the present invention might find use;

FIG. 2 shows a more detailed block diagram of typical endnode in the network of FIG. 1 at which point packets may enter the network to be forwarded along the route to a destination for each packet, and in which network access mechanisms in accordance with the present invention might be implemented;

FIG. 3 shows a general block diagram of an enhanced rate-based network congestion control system in accordance

of the end nodes 12 of FIG. 1. The endnode control circuit of FIG. 2 comprises a high speed packet switching fabric 33 onto which packets arriving at the node are entered. Such packets arrive over transmission links from network nodes of the network, such as links M-O of FIG. 1 via transmission interfaces 34, 35 or 36, or are originated locally via local user interfaces 30, 31 or 32 from user applications 40, 41, ..., 42. Switching fabric 33, under the control of connection management facilities 44, connects each of the incoming data packets to the appropriate one of the outgoing transmission link interfaces 34-36 or to the appropriate one of the local user interfaces 30-32, all in accordance with well known packet network operations. Indeed, network management control messages are also launched on, and received from, the packet network in the same fashion as data packets. That is, each network packet, data or control message, transmitted on the network of FIG. 1 can be routed by way of switching fabric 30, as shown in FIG. 2.

Routes or paths through the network of FIG. 1 are calculated to satisfy the Quality of Service (QoS) parameters determined to be necessary to adequately transmit a particular data stream, as taught in the afore-mentioned patent 5,233,604. These Quality of Service parameters include such things as throughput (bandwidth), latency (path delay) and jitter (latency variations). Assuming that only one class of traffic requires preselected QoS parameters, called reserved traffic, and that at least one other class of traffic can be transmitted on a "best efforts" basis, the maximum use of the network occurs when any available capacity not currently used by the reserved traffic is filled with best effort traffic. Since the reserved traffic must be transmitted in a timely fashion, it is necessary to control the access of the best efforts traffic so as not to interfere with the reserved traffic. Since, due to changes in traffic loading or outages, congestion occurs in the packet network of FIG. 1, it is necessary to restrict the access of best effort traffic to the network in such a way as to reduce the congestion in the network and allow the reserved traffic to flow unimpeded. One known technique for controlling access to the network is to estimate the congestion in the network and to adjust the input to the network so as to reduce such congestion. In accordance with one well known congestion estimation technique, taught in R.-F. Chang et al. patent 5,367,523, granted November 22, 1994 and in "Adaptive Admission Congestion Control," by Z. Haas, published in *ACM SIGCOMM Computer Communication Review*, pages 58-76, 1991, a sample test packet is transmitted from the source of each data stream to the destination for that data stream. From time stamps carried by the test packets, the receiving station estimates congestion by identifying changes in the transit times of the test packets. The present invention is an improvement on these latter systems using test packets to estimate the congestion in the system of FIG. 1. More particularly, the present invention attempts to estimate the actual level of congestion in the network, rather than relying on relative congestion, i.e., changes in congestion, during successive test sample transmissions. This approach can be better seen from the congestion estimation system in the block diagram of FIG. 3.

Returning to FIG. 2, connection management facilities 44 utilize the contents of network topology data base 45 to calculate reserved and best efforts paths through the network of FIG. 1 in response to requests from users 40-42. A sample packet processor 37 generates sample test packets for transmission from the end node of FIG. 2 as well as providing the received sample packet processing necessary for test sample packets originating at other end nodes and intended for the end node of FIG. 2. Using the results of this received sample processing, network admissions processor 43 generates control signals to control the access of best effort signals from user applications 40-42 in user application adapters 30-32, respectively.

Referring then to FIG. 3, there is shown a general block diagram of the test sample packet processing taking place in blocks 37 and 43 of FIG. 2. More particularly, reserved data source 17 corresponds to one of the user applications 40-42 and delivers reserved data packets to network entry buffer 29 for delivery to network 23, corresponding to the network of FIG. 1. This reserved data traffic is delivered through network 23 to reserved user 18 at a remote end node of network 23. An unreserved data packet source 20 corresponds to another one of user applications 40-42 and generates non-reserved (best efforts) data packets for transmission on network 23, corresponding to the network of FIG. 1. An adaptive admissions control circuit 21 controls the rate at which unreserved data packets from unreserved data source 20 are applied to network entry buffer 22, which, in turn, supplies the sample packets to network 23. The adaptive admissions control circuit 21 is of a type well known in the art such as that disclosed in H. Ahmadi et al. patent 5,311,513, granted May 10, 1994, and will not be further described here. A test sample packet generator 27 generates test sample packets which are applied through admission control circuit 21 to network entry buffer 22 for transmission on network 23.

At the remote destination end node, the data packets from sources 17 and 20 are delivered to reserved/unreserved users 18. At the remote location, a sample packet analyzer 24 receives the test sample packets generated in sample packet generator 27, processes these samples to obtain an estimate of the absolute value of the current state of congestion in the path between data packet source 20 and sample packet analyzer 24. In accordance with the present invention, the state of congestion is approximated by the current length of the queuing delay in network 23. In response to this estimate of congestion, sample acknowledgment generator 25 generates an acknowledgment of the receipt of each sample packet and transmits this acknowledgment back over network 23 containing a coded estimate of the current queuing delay on this connection through network 23. This acknowledgment is delivered through network 23 to sample acknowledgment processor 26 at the source end node of the original sample packet. This sample acknowl-

delay is made up of the transmission delays in all of the links in the path, the propagation delay, the processing delay, and the queuing delays accumulated at the nodes in the path, to accommodate the traffic flowing through the nodes. The minimum delay measured at the receiving station, then, is the accumulated total delay, assuming that no queuing delay is necessary at intermediate nodes. It is the queuing delay, and the queuing delay alone, that is a measure of the congestion. In order to allow the network to reach an uncongested state before excessive adaptive changes, and thus to get a good approximation of the queuing delay, it is important that the initial rate of sending sampling cells be set relatively low.

To illustrate the rapid convergence of the queuing delay estimation, FIG. 4 illustrates the transmission of the first three successive sampling cells on a particular connection. As can be seen in FIG. 4, the network delays 52, 53 and 54, are assumed to be monotonically decreasing. Note that the error ( $ER$ ) at each successive sampling cell is reduced and never grows larger for any possible future sampling cell delays. That is, the queuing delay estimate error goes to zero whenever the queuing delay itself goes to zero, at any time during the connection.

The fourth element of the heuristic algorithm of the present invention is the division of the estimated delays into four categories representing four different categories of congestion, called congestion states (CSs):

1) **Network is free.** Virtually no delay has been detected. All non-reserved connections are allowed to increase their transmission rates. This congestion state is coded "01," i.e., CS=01.

2) **Desired Network Condition.** A desired network utilization level has been achieved. All connections should maintain their current rates. This congestion state is coded "00."

3) **Slight Congestion.** A small amount of congestion has been detected in the network. All connections should take a small reduction from their current rates of transmission. This congestion state is coded "10."

4) **Serious Congestion.** A serious congestion condition has been detected in the network. All connections should take a large reduction from their current transmission rates. This congestion state is coded "11."

These different congestion conditions are coded into the sampling packet acknowledgment signal from the receiving station to appropriately control the rate changer at the sending station.

The fifth and last element of the heuristic congestion control algorithm of the present invention is the strategy used to adjust the non-reserved transmission rate when a particular level of congestion is detected. In order to ensure efficiency, stability and fairness, a large change in the allowed rate is permitted when the allowed rate is low (to increase efficiency), but only a small change in the allowed rate is permitted when the allowed rate is already high (to increase stability). A logarithmic relationship between rate changes and congestion (delay) changes is used in the heuristic algorithm of the present invention. In order to ensure fairness and efficiency, the heuristic algorithm takes the following further steps:

1) Modify the allowed rate, based on a logarithmic function such as:

$$X(t+1) = Cx1 \ln\left(\exp\left(\frac{X(t)}{C}\right) + \Delta\right), \quad (2)$$

where  $X(t)$  and  $X(t+1)$  are the allowed rates at times  $t$  and  $(t+1)$ , respectively,

$C$  is a scaling factor to be selected heuristically,

$\Delta$  is a constant which, for convenience, is set to 1.

2) Determine the current fair share of the remaining bandwidth, at the starting point after the available rate is determined, which is available for the new best efforts connection.

3) Scale all of the non-reserved transmission rates, if necessary, to ensure fairness in sharing and efficient use of the total available bandwidth (available after all reserved traffic is served).

If  $r$  is the number of reserved connections on the link, the remaining bandwidth  $C_A$  available for best efforts services is given by

but values of 0.9 and 0.5, respectively, have been found to be usable.

The second step of the rate change policy involves assuming that an even share of the available bandwidth has been achieved if the variable STATE oscillates within a predefined "Damping Region". This Damping Region (illustrated in FIG. 5) is initialized as  $(L, L + DR)$  where  $L$  is the lower limit (e.g., curve 62) and  $DR$  is the size of the damping region (e.g., the region between curve 60 and 62), where  $L$  and  $DR$  are determined heuristically, observing actual stability values. In FIG. 5, it is assumed that  $DR$  is equal to ten. As previously noted, curves 60, 61 and 62 are obtained by varying the constant  $C$  in equation (5) to obtain the various rate values. The middle curve 61 is assumed to be the initial rate table values, and curve 60 represents the rate table values after the rates have moved up five times. Curve 62 represents the rate table values after the rates have move down five times. For illustrative purposes, these values represent an initial Damping Region between curves 60 and 62. If STATE stays within the Damping Region, the table  $R$  is considered to be appropriate for the current network conditions.

If STATE grows outside of the Damping Region, due to successive increase acknowledgments, the third step in the rate change policy is used. In this case, the entire table is increased by multiplying each value by a predefined factor greater than one, defined as moving the damping region up by one step, i.e.,  $L = L + 1$ . If this occurs, the value of the variable STATE is reset to its initial index value. On the other hand, if STATE never reaches the value  $L$ , the table  $R$  is considered to be too high and the entire table  $R$  is reduced by multiplying each value in the table by a predefined factor less than one, defined as moving the damping region down by one step, i.e.,  $L = L - 1$ . The value of the variable STATE is then reset to its initial index value. The multiplicative modification of the table  $R$  preserves the logarithmic characteristic of the rate. Fairness between the different best efforts connections may be lost since they would all be using different rate tables due, possibly, to different start up times. The linear movement of the Damping Region, and the resetting of the value of STATE, tends to favor the connections using the lower valued tables versus those using higher value tables, thus minimizing the fairness divergence. This entire rate change policy is summarized in the pseudo-code given in the APPENDIX.

Rather than starting the initial transmission rate at zero, requiring a long start-up time to come to an appropriate rate, equation (4) is used when each new best efforts connection is started to set the initial rate at a value satisfying the basic fairness criterion. Rates for all connections can also be reset when the Damping Region is moved. It is noted, however, that fairness of access in the sense defined in equation (4) does not necessarily maximize the total network throughput since some traffic sources (e.g., file servers or routers) inherently use more bandwidth than other traffic sources. It may therefore be desirable to allocate bandwidth to the various traffic sources in a weighted manner, reflecting their average bandwidth requirements. This is sometimes called proportional fairness.

To allow for unequal allocation of available bandwidth to best efforts connections, each best effort traffic source is allocated an initial rate that is proportional to its need. The fairness function for such a proportional fairness policy can be defined as:

$$S_j = \frac{MCR_j}{\sum_{i=1}^n MCR_i}, \text{ where } (6)$$

$MCR_j$  is the minimum cell rate guaranteed by the network to best effort connection  $j$ .

Turning now to FIG. 6, there is shown a typical rate table consisting of essentially four columns. The first column is the STATE or index value which is not actually in the table, but assists in understanding the operation of the table. The second column is the transmission rate. Successive transmission rates in the table of FIG. 6 are related to each other logarithmically as suggested in FIG. 5. The third column of the rate table of FIG. 5 are the next STATE or index value to go to for the slight congestion condition while the fourth column are the next STATE or index value to go to for the severe congestion condition. That is, the next rates used under slight congestion conditions are different from (less than) the next rates used for severe congestion conditions. This allows the rate to move down faster in severe congestion conditions than in slight congestion conditions, all as described above to insure stability in the unreserved data transmission rate. This process can be better seen in connection with the following description of the flow charts of FIGS. 7 through 9.

In FIG. 7 there is shown a flow chart of the procedures taking place in the source node for an unreserved, best efforts connection to launch new sample packets needed to measure network congestion. Starting in start box 50, box 51 is entered where the connection variables are initialized. These variables include:

The Round Trip Time (RTT) estimate is the initial value of the round trip time required to transmit a packet from

In FIG. 8 there is shown a flow chart of the acknowledgment processing taking place at the source of the unreserved traffic being monitored and adaptively controlled. Starting at the upper left of FIG. 8, the flow chart splits into four branches 70, 71, 72 and 73, corresponding to the four congestion states coded into the acknowledgment, as will be discussed in connection with FIG. 9. These four congestion states correspond to codes "00" (congestion acceptable as within the desired range), "01" (congestion below target levels), "10" (congestion slight) and "11" (congestion serious). In branch 70 (CS=00), nothing is done since the rate STATE is acceptable. In branch 71, STATE is incremented by one to increase the data transmission rate to the next value in the table of FIG. 6. In branch 72 (CS=10), the rate state is set at a value equal to the product of a constant DE1 (noted above as a small reduction factor, such as 0.9) and the minimum of the current value of STATE and the maximum permissible value of STATE (S). This results in a value for STATE which is reduced from the current STATE by the small reduction factor DE1. In branch 73 (CS=11), the rate state is set to a value equal to the product of a constant DE2 (noted above as a large reduction factor, such as 0.4) and the minimum of the current value of STATE and the maximum permissible value of STATE (S). This results in a value for STATE which reduced from the current value of STATE by a large reduction factor DE2. All of branches 70, 71, 72 and 73 are entered into box 74 where the value of STATE is set to the value calculated in branches 70-73 or to "1," whichever is higher. That is, the reduction in the value of STATE is not allowed to fall below the value "1."

After the new value of STATE is calculated, as described above, decision box 75 is entered to determine the range of values into which the calculated new value of STATE falls. As can be best seen in FIG. 5, the possible range of values of the rate  $R(i)$  is divided up into a damping region, between curves 60 and 62, a move-down region, above curve 60, and a move-up region, below curve 62. The move-up region is detected in decision box 75 where the value of STATE is less than L. Box 76 is then entered where the count UD is incremented by one. It will be recalled that UD is a count of the number of data cells transmitted since the value of STATE was below L. Decision box 77 is then entered to determine whether the value of UD has exceeded the value UD\_TH. UD\_TH is a threshold which, when exceeded, causes the desired damping region to be scaled down so as to encompass the recent values of STATE. In box 78, this rescaling is accomplished by dividing each rate value in the second column of the rate table (FIG. 6) by a fixed value ("ADJ"). This reduction in the rates moves all of the rate values down to span the newly calculated STATE values. In box 79, the value of L is incremented by one to properly represent the new lower level of the damping region (FIG. 5). The process then returns to FIG. 7 to trigger the transmission of a new sample cell and to utilize the new rate value for future data transmissions.

Returning to decision box 75, if the value of STATE is equal to L, that is, the STATE value is equal to the lower boundary of the damping region, box 81 is entered where the value of UD is reset to zero, indicating that the number of sampling cells below the damping region must be reset. If the value of STATE is greater than the sum of L and DR, that is, if the value of STATE is above the damping region, box 82 is entered to rescale the values of the data rates upward, thereby accommodating the new value of STATE. The rescaling is done with the same factor ADJ used in box 78, but used multiplicatively instead of by division to increase the values of the data rates in the second column of the table in FIG. 6. After rescaling the data rates in box 82, box 83 is entered where the value of STATE is set to the maximum value S and the value of L is incremented by one. The process then returns to FIG. 7 to send a new sample cell and to use the new rate.

FIG. 9 shows a flow chart of the processing done at the remote destination node in boxes 24 and 25 of FIG. 3. Starting at start box 90, box 91 is entered where the variables are initialized at the remote destination node. For example, the values of Q1 and Q2, the previous (Q1) and the current (Q2) network delays, are initialized at the value of the calculated round trip transit time calculated from the data in the topology data base 45 (FIG. 2) located at that node. The sample cell transmit time and receipt times are both initialized at zero. Box 92 is then entered to await the arrival of the next sample cell. The first sample cell over any connection includes the values of D1, D2 and D3, calculated in box 51 of FIG. 7, to be used at the remote terminal of FIG. 9. When a sample cell arrives, box 93 is entered where the delay is estimated, using equation (1). Once a delay estimate is available from box 93, this estimate is successively applied to decision boxes 94, 95, 96 and 97 to determine the congestion status of the network as described above. Decision box 94, for example, determines whether the delay estimate  $Q_2$  is less than or equal to D1. If so, the congestion status is "01" indicating that the congestion in the network is less than the preferred level and hence the unreserved data rate should be increased. An acknowledgment cell bearing this congestion status code is transmitted from block 98 of FIG. 9.

Decision box 95, on the other hand, determines whether the estimated delay is between D1 and D2, indicating that the network delay is in the desired range. The congestion status code "00" is encoded in the acknowledgment cell generated in box 98. If the estimated delay is between D2 and D3, as determined in decision box 96, indicating that slight congestion is occurring in the network, the congestion code "10" is encoded in the acknowledgment cell generated in box 98. Finally, if the estimated delay is greater than D3, as determined in decision box 97, indicating that serious congestion is occurring in the network, the congestion code "11" is encoded in the acknowledgment cell generated in box 98. As noted in connection with FIG. 8, these congestion states are used to process the STATE variable at the source node. Once the acknowledgment cell is transmitted in box 98, box 92 is re-entered to wait for the next sample cell.

## APPENDIX

## Rate Change Pseudo Code

## PROCEDURE RATE\_CHANGE\_POLICY

ADJ = table adjustment factor where  $ADJ > 1$ .

DE1, DE2 = decrease factors 1 and 2 where  $DE2 > DE1$ .

DR = size of the damping region.

S = size of the rate table.

STATE = index pointing to current allowed rate in rate table  
R.

L = lower bound of damping region.

UD = number of sampling cells sent while STATE is below L.

UD\_TH = threshold value for UD.

$X(t)$  = allowed rate at time  $t$ .

## Initialization

L = S; initial level

UD = 0;

UD\_TH = 2 x L; threshold for UD to take action

STATE = 1

$X(t) = R(STATE)$ ; initial rate

## Case of ACKNOWLEDGMENT

INCREASE: STATE = STATE + 1

NO CHANGE: Do nothing

SLIGHT CONGESTION: STATE = [DE1 x min(STATE, S)]

SERIOUS CONGESTION: STATE = [DE2 x min(STATE, S)]

endcase; ACKNOWLEDGMENT

## Case of STATE

> L: UD = UD + 1

if UD > UD\_TH

$R(i) = R(i)/ADJ$ , for  $i = 1, 2, \dots, S$

L = L - 1

endif

< L + DR:  $R(i) = R(i) \times ADJ$ , for  $i = 1, 2, \dots, S$

STATE = S,

L = L + 1

6. The packet transmission network according to claim 1 further comprising

means for reducing said rate of said best efforts data by a first smaller value when said rate is at a lower value,  
and

means for reducing said rate of said best efforts data by a second value greater than said first value when  
said rate is at a higher value than said lower value.

7. The packet transmission network according to claim 1 further comprising

means for initializing said rate of said best efforts data at the fair share value of the currently available bandwidth  
in said network.

8. A method for transmitting maximum data on a packet transmission network comprising the steps of

interconnecting a plurality of transmission nodes by transmission links,

transmitting data streams from a plurality of user applications on said network, said data streams including  
reserved path data streams requiring guaranteed levels of transmission quality and best efforts data streams  
requiring only the best efforts available levels of transmission quality,

selecting reserved data paths through said network between pairs of said user applications wherein said re-  
served data paths have said guaranteed levels of service quality,

transmitting best efforts data on said network during those times when said reserved data paths are partially  
idle,

controlling the access of said best efforts data to said network in order to prevent interference with data on  
said reserved data paths, said step of controlling the access of said best efforts data comprising the steps of

detecting the level of congestion in said network,

controlling by means of a look-up table the rate of said best efforts data in response to said level of congestion,  
and

multiplicatively changing the values in said table driven means in response to excessive changes in said rate  
of said best efforts data.

9. The method according to claim 8 wherein said step of detecting congestion in said network comprises the steps of  
estimating the absolute value of the queuing delay in said network.

10. The method according to claim 8 further comprising the step of  
inserting a spacer between successive cells of said best efforts data.

11. The method according to claim 10 wherein said step of inserting a spacer comprises the step of  
selecting said spacer length in response to a time-out function, to an acknowledgment, or to the transmission of  
a minimum burst of said best efforts traffic, whichever occurs first.

12. The method according to claim 8 wherein said step of determining the congestion in said network comprises the  
steps of

transmitting a series of test samples through said network each having a time stamp containing the time of  
launch,

comparing successive test samples to determine total delay through said network, and

determining the minimum value of said total delay through said network during the duration of a connection.

13. The method according to claim 8 further comprising the steps of



FIG. 1

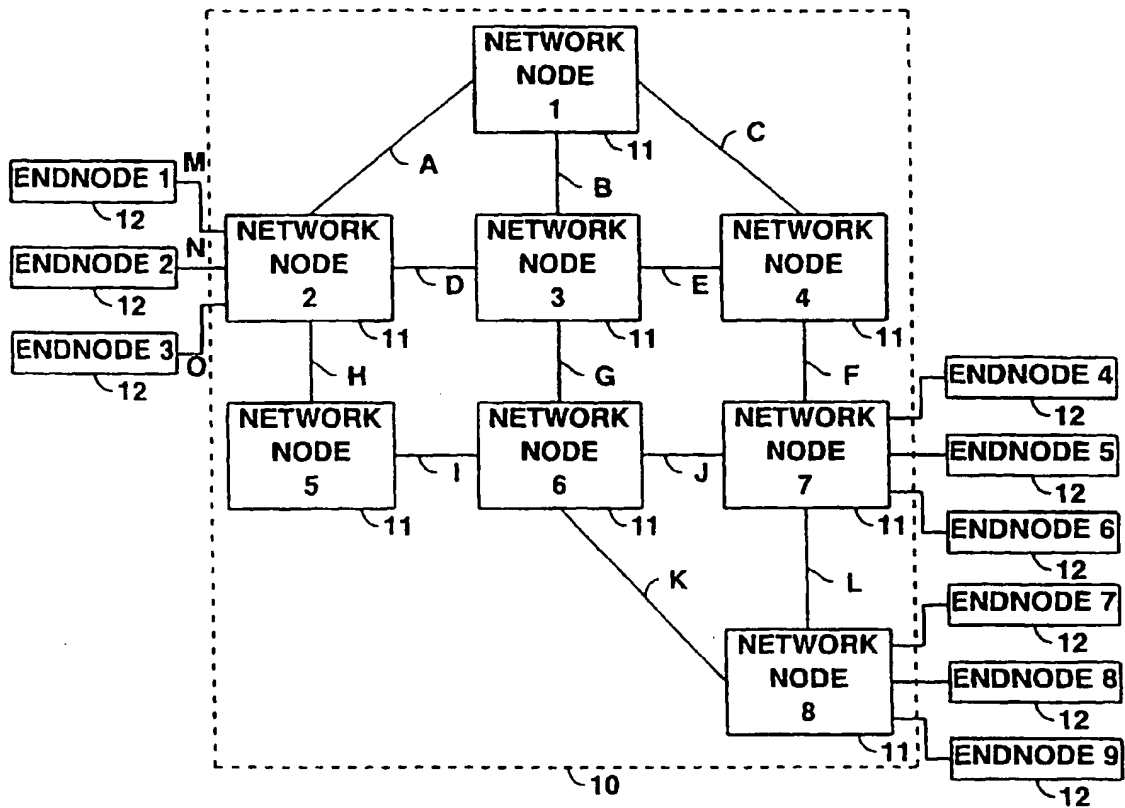


FIG. 3

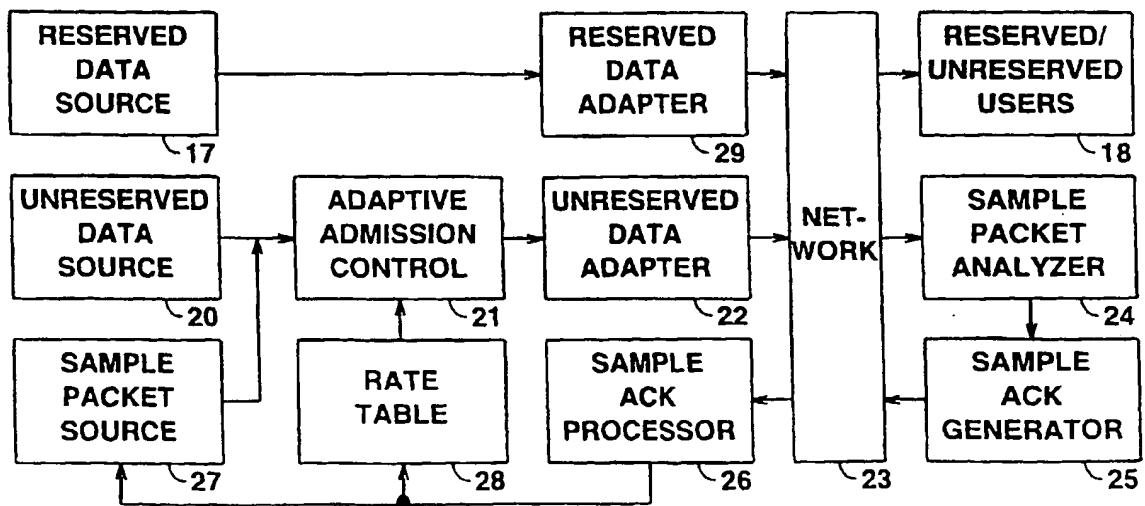


FIG. 4

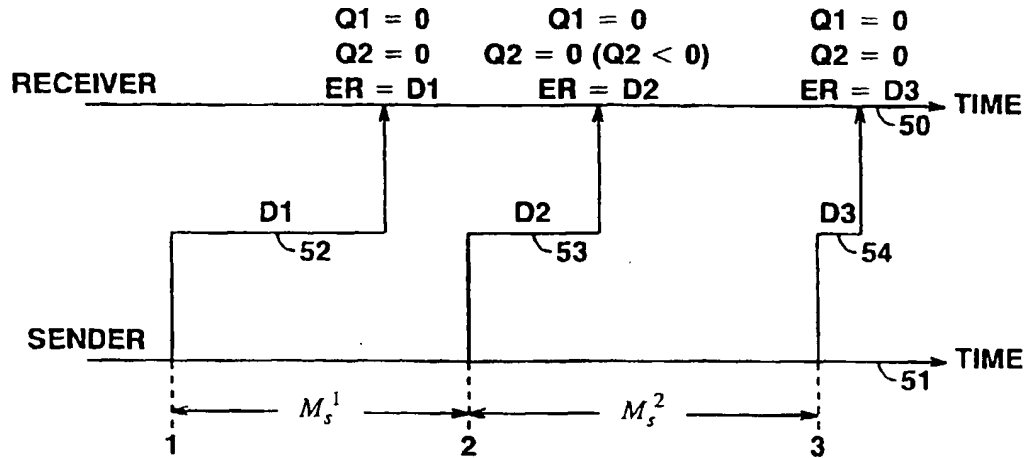


FIG. 5

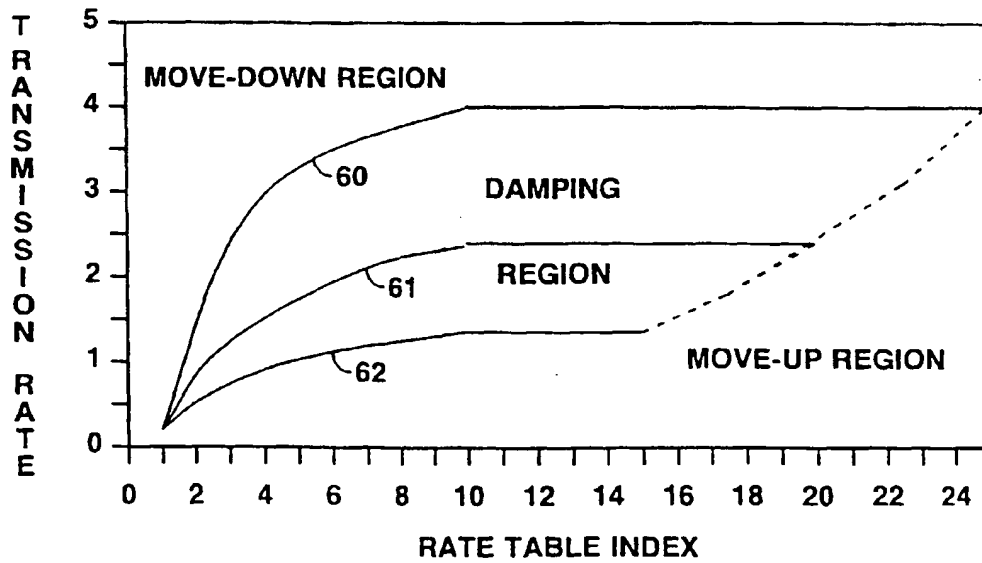
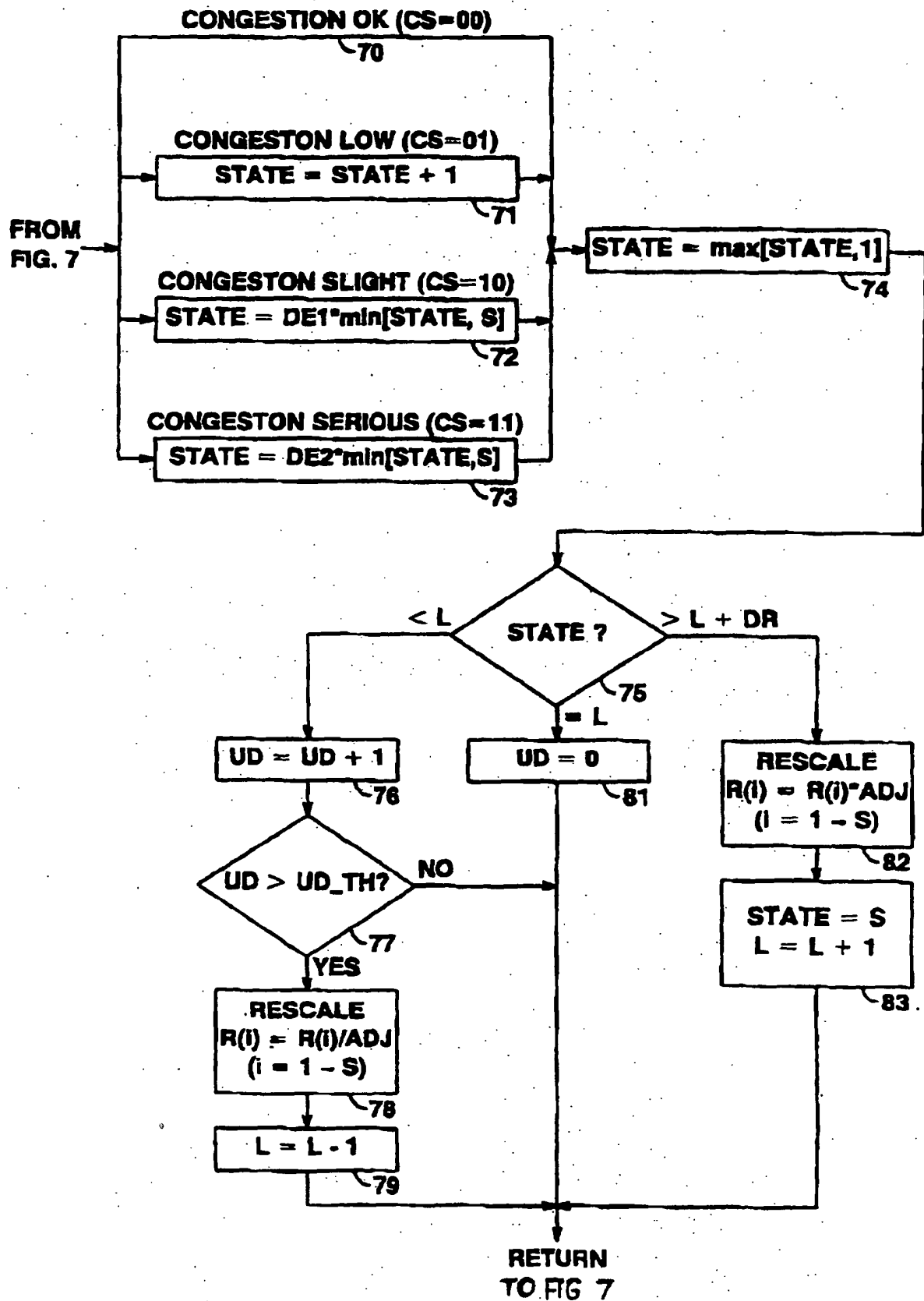


FIG. 8



(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

**EP 0 827 307 A3**

(12)

**EUROPEAN PATENT APPLICATION**

(88) Date of publication A3:  
30.08.2000 Bulletin 2000/35

(51) Int Cl.7: **H04L 12/56**

(43) Date of publication A2:  
04.03.1998 Bulletin 1998/10

(21) Application number: **97480049.2**(22) Date of filing: **19.08.1997**

(84) Designated Contracting States:  
**AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC  
NL PT SE**  
Designated Extension States:  
**AL LT LV RO SI**

- Onvural, Raif O.  
Cary, NC 27513 (US)
- Huynh, Lap T.  
Apex, NC 27502 (US)
- Gun, Levent  
Hopkinton, MA 01748 (US)
- Van Vu, Ken  
Cary, NC 27513 (US)
- Ling, Bouchung  
Cary, NC 27513 (US)

(30) Priority: **22.08.1996 US 701624**

(71) Applicant: **INTERNATIONAL BUSINESS  
MACHINES CORPORATION**  
Armonk, NY 10504 (US)

(74) Representative: **de Pena, Alain**  
**Compagnie IBM France**  
**Département de Propriété Intellectuelle**  
**06610 La Gaude (FR)**

(72) Inventors:

- Marin, Gerald Arnold  
Chapel Hill, NC 27514 (US)

(54) **Adaptive rate-based congestion control in packet networks**

(57) An enhanced adaptive rate-based congestion control system for packet transmission networks uses the absolute rather than the relative network queuing delay as the measure of congestion in the network. Other features of the congestion control system include test transmissions only after a predetermined minimum time, after the receipt of an acknowledgment from the previous test, or transmission of a minimum data burst,

whichever takes longest. The congestion control system also provides a small reduction in rate at low rates and a large reduction in rates at high rates. A logarithmic rate control function provides this capability. Rate damping is provided by changing all of the values in a rate look-up tables in response to excessive rate variations. Finally, the fair share of the available bandwidth is used as the starting point for rates at start-up or when a pre-defined rate damping region is exited.

**FIG. 1**